Search Engine Methodology
What's in a Search Engine?

To effectively optimize for search engines and to better understand what's really happening, there is value in knowing how modern search algorithms work. This article will walk through the creation of a hypothetical search engine, and will show how this impacts search engine optimization.

Step One: Make a List of URLs and Crawl Them

Before anything can be done, a list of URLs needs to be retrieved to initially crawl. The most popular option for this is to load the URLs in the DMOZ database. These aren't the only sites that will be crawled. The pages linked to by sites in the DMOZ directory are also crawled since the crawler follows the links. It certainly helps to be in DMOZ, especially if you don't have enough links from other sites to be sure that you'll be sufficiently crawled.

Now, a group of computers are set up to download all of the pages on the list. These are called the "crawlers." They will also look at the links on those pages, and crawl those URLs as well (the crawlers will continue following links until their hard drives are full).

Step Two: Analyze the Pages

The crawlers now go through each page and look at their content.

First, the crawler makes a table with every unique word on the page. It gives "points" to each word based on how many times it's used on the page, and words in bold, in the title, in meta tags, or in headers are given extra points.

Word	Points
shoes	145
athletic	78
sneakers	34
sandals	12
(etc.)	

This means that you should use the most important words more often in your text. However, using a word too often will mark your page as

being spam, which will cause the crawler to delete your site from its database.

It then creates a percentage of the frequency of each term:

Word	Points	Percentage
shoes	145	5.80%
athletic	78	3.12%
sneakers	34	1.36%
sandals	12	0.48%
(etc.)		

Usually, the percentages are stored in the database and not the actual points, though longer pages may be given a slight advantage later on. As a result, adding a lot of unnecessary text that uses one term a lot will raise your percentage for that term, but will also lower the percentage for other terms.

More advanced engines will also cross-reference each word to other major words based on where they are relative to each other. (Words appearing next to each other are given more points here.) So, for example:

Word	shoes	athletic	sneakers	sandals
shoes	-	20	12	7
athletic	20	-	11	4
sneakers	12	11	-	5
sandals	7	4	5	-

As a result, the placement of words relative to each other *does* matter. This is why targeting phrases is usually better than targeting a variety of single words.

Calculate Link Popularity

The crawlers now take their lists of the URLs that each page links to and combine them. So for each page there is now a list of the links on it, as well as the text of each link. The list is then reversed, so that instead of showing the links on each page, it shows for each page the sites that link to it.

Some search engines stop here and simply store the number of links pointing to a given page, but Google takes it a little further.

For every page in its database, Google gives it "points" based on how many links are going to it--just like any other search engine. Then, it re-calculates the number of links pointing to each page, but gives more points to links that had a higher point-value themselves in the first count. It then repeats the process about 100 times, each time making the points more accurate. So:

1. Points are assigned based on the number of links going to a page.

2. Points are calculated again, but pages get more points if the links going to a page had more points in the last step. (Because Yahoo! had a lot of links going to it in the first step, a link from Yahoo! would now be more valuable.)

3. The original point values are thrown out and are replaced with the points just calculated. Now, the points are re-calculated again, this time considering the points from Step 2 instead of Step 3. This is repeated approximately 100 times, and every time the points become more accurate (because it considers further down the line where links are coming from).

Now, Google takes the point values--which could be extraordinarily large--and converts them to a PageRank, which is on a scale of 0 to 10. However, it does *not* simply convert, for example, 1,000 to 1 and 2,000 to 2. The scale is logarithmic, which means that higher PageRanks require much more points.

WebmasterGoodies has an approximation of what the ranges most likely are--look at the first three columns. The actual ranges aren't available to the public, but the ranges on that site are believed to be fairly close. Obviously, a logarithmic scale makes a difference: PR1 requires 6 to 30 "points," while PR10 requires more than 25 million points.

Now What?

Search engines put the databases into a specialized format, and then write the search software.

When a search is made, every site containing the relevant terms is pulled up. The ranking is based on a combination of the points for each

relevant term, the site's link popularity (PageRank), and other smaller factors. Each engine weighs these differently.

You should now have a better understanding of what's happening under the hood of the search engine, and this should help in optimizing your pages as well.